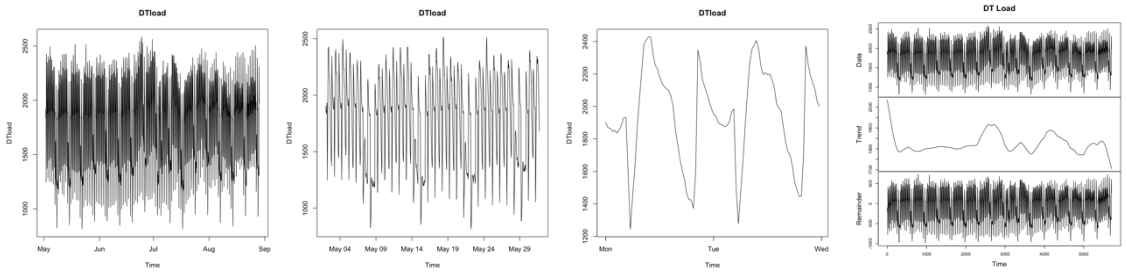
Steve Depp
413 – 55

Assignment 10
2 September 2019

Random Forest

1.1 Perform EDA.



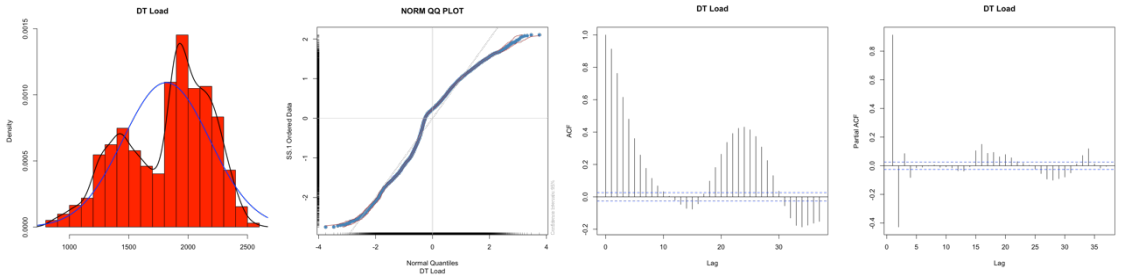
Data span May – Aug 2016 or 5712 points. Plot on **left above** shows some dip. Focusing on the first month, a general cycle in load is visible: 2 spikes, one in late June and another around July 4th followed by load decline in mid-July. Would guess these could be predictable based on external data: temperature and holiday. Plot in the **middle left above** represents the 1st month with generally lower lows at the end of the month but overall stable, and a discernable seasonality that appears to be once a week lower Load on weekends: there are 31 days in May and so there are more than 4 full 7-day seasons in May. Plot on **middle right above** shows daily seasonal pattern, higher in the daytime and lower at night. Strip out the trend from the distribution, as shown on the **right above** and the general visual mean appears steadier though one can make out a broader range in values on the left center vs extreme left and on right side of plot.

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
X1	5712	1816.544	365.0802	1901.366	1835.286	814.5063	2584.202	1769.696	-0.4576675	-0.6973716

Univariate statistics **above** suggest a distribution with negative skew and kurtosis. Skew is supported by mean < median. Appears the center of the distribution is negatively skewed as well by the trimmed < median. Tests allow rejection of hypotheses of zero mean, skew, kurtosis.

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
X1	10	362.3916	27.15867	367.5848	362.8929	322.0317	398.7412	76.70949	-0.2644712	-1.608572

The **above** represent the univariates of the standard deviation of Load, consistent with the generally wider Load ranges visible in the above plots in May and June.

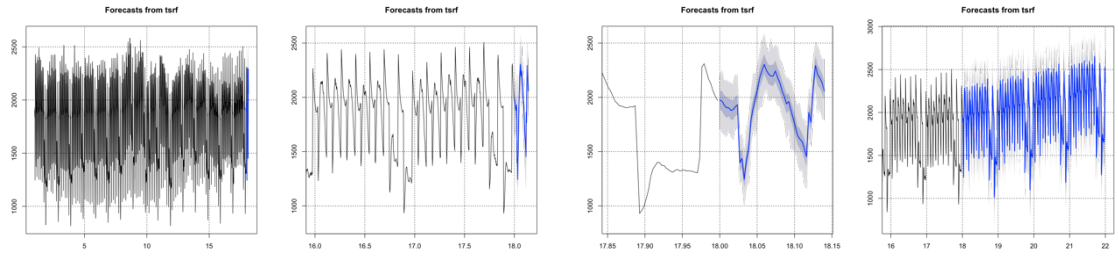


The **above** histogram represents bimodal distribution derived from regular seasonality. QQ generally normal and platykurtic: boxed +/- 2 SDs versus normal. The ACF and PACF represent serial correlation in wave fashion that might suggest Fourier transformation is useful for modeling.

1.2 Model is built with 500 trees, 3 candidates at each split, node size = 5, KK=2 for double Fourier seasonality, and applied to the detrended series (since RF cannot predict trends).

1.3 – 1.4

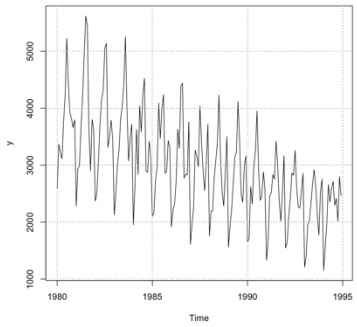
	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
18.00000	1974.399	1886.291	2064.571	1713.111	2206.023
18.00298	1966.049	1873.259	2057.924	1681.914	2247.864
18.00995	1929.455	1837.433	2022.429	1653.026	2154.558
18.00893	1905.107	1809.891	2001.477	1638.560	2162.653
18.01190	1903.405	1826.293	1993.368	1698.823	2142.716
18.01488	1880.496	1796.718	1976.907	1634.474	2129.253



The **above** table shows the first few forecast predicted DT Load values and their low and high intervals. The **above** plots show the forecasts in blue with grey intervals generated from bootstrap resampling of the RF model. Left 2 plots are difficult to analyze. On the **right middle above**, the medians appear to capture daily seasons correctly, with peaks in the middle and do capture weekend dips as shown on **extreme right above**. Somehow a trend is incorporated in these data, possibly some portion of the AR model in the code did not capture all.

2.0 Neural Network

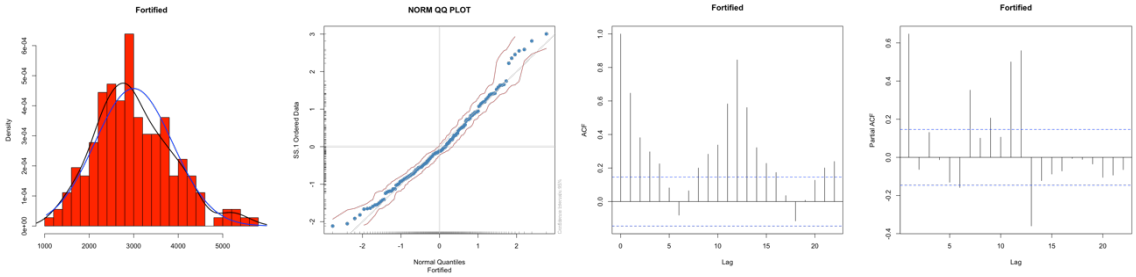
2.1 Use the EDA from the earlier assignment that used the Australian wine data



	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
X1	180	2998.544	872.0442	2894.5	2958.66	1154	5618	4464	0.4939306	0.1519758

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
X1	10	699.5886	137.4697	691.0375	697.0899	464.5649	954.6022	490.0373	0.1214319	-0.8220256

Principal features of the plotted series **above** is a strong negative trend that finished 2 years ago, and which has smaller ranges on the right than left sides: The level has fallen from ~4000 mean annual value in Jan 1980 to ~2000 through about 1993. Fitting that the overall mean as shown in the 1st univariate table **above** is 2998.5. Variation has declined proportionately with level from variations of 3000 per year to less than 2000. In the last 2 years, variation has compressed significantly to ~750 and the level is in the top end of the 2-year range. Variation in standard deviation is quantified in the **2nd univariate table above** which shows that standard deviation ranged from 955 max (presumably on the left of the plot) to 465 (probably on the right of the plot). It is noteworthy that the starting bottles and finishing bottles are roughly similar ~2500. Lastly in term of seasonal variation, over 15 years one can count 15 spikes up and 15 spikes down with some intra year variation since 1983 that is about half the length of those spikes up and down. The spikes from high to low have exceeded the mean level of wine sales in some circumstances and are therefore significant.



Tests allow rejection of hypothesis of symmetry at the 99% confidence level, but not the hypothesis of excess kurtosis. The positive numerical skew in the **table above** is depicted in the histogram **above.** Otherwise, the QQ and histogram appear normal. The ACF demonstrates some form of wave of serial correlation, but I am not skilled enough to know whether this qualifies as sin/cos fitting for a Fourier transform. PACF demonstrates some partial correlation at the 12-month lag point. So possibly there is annual seasonality to explore. Box Ljung tests allow us to reject the null of no serial correlation. The ADF test of unit roots allows rejection of the null of no unit roots for "constant with trend" and the KPSS test suggests a 1st difference that might be consistent with the strong serial correlation in the PACF at lag 1. Possibly therefore there is a general monthly trend down in wine sales offset by an annual seasonal wave.

```
a 12-6-1 network with 85 weights
options were - linear output units
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1
   0.56    0.00    0.34   -0.85    0.81   -0.44   -0.30   -0.24    0.83   -1.98
i10->h1 i11->h1 i12->h1
   2.49   -3.71   -0.37
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2  i9->h2
  -1.04   -4.59    2.23    2.01    0.12    1.58   -0.84    0.85   -0.43   -4.01
i10->h2 i11->h2 i12->h2
  -0.36   -0.26    0.83
  b->h3  i1->h3  i2->h3  i3->h3  i4->h3  i5->h3  i6->h3  i7->h3  i8->h3  i9->h3
   0.01   -0.69    0.56    0.73   -0.95    0.66    0.21    0.17   -0.97    0.15
i10->h3 i11->h3 i12->h3
  -1.08    0.63   -0.20
  b->h4  i1->h4  i2->h4  i3->h4  i4->h4  i5->h4  i6->h4  i7->h4  i8->h4  i9->h4
  -0.33   -0.64   -1.40   -0.87    0.34    0.00    2.80   -2.29   -0.92   -1.88
i10->h4 i11->h4 i12->h4
   2.51   -1.01   -0.06
  b->h5  i1->h5  i2->h5  i3->h5  i4->h5  i5->h5  i6->h5  i7->h5  i8->h5  i9->h5
   1.04    2.58    0.16    1.39   -0.91    0.53   -2.82    2.76    0.50    2.57
i10->h5 i11->h5 i12->h5
  -3.34    2.49    0.33
  b->h6  i1->h6  i2->h6  i3->h6  i4->h6  i5->h6  i6->h6  i7->h6  i8->h6  i9->h6
   1.67    1.45   -0.77    1.64   -2.14    1.81   -0.01    0.54   -2.47    0.68
i10->h6 i11->h6 i12->h6
  -0.75   -0.16    1.70
 b->o h1->o h2->o h3->o h4->o h5->o h6->o
 3.29 -2.05  0.76 -3.48 -2.31 -2.50  2.56
```

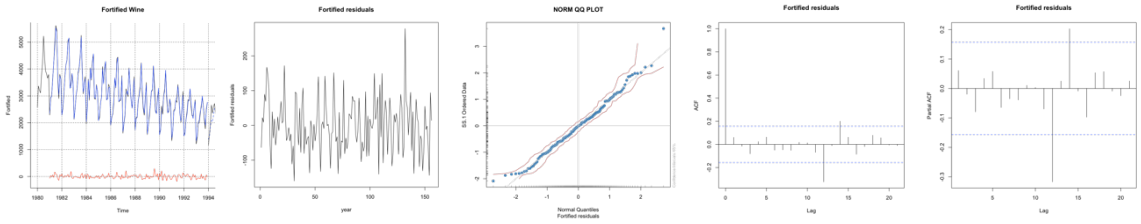
```
Series: x.tr
Model:  NNAR(11,1,6)[12]
Call:   nnetar(y = x.tr, p = 11)

Average of 20 networks, each of which is
a 12-6-1 network with 85 weights
options were - linear output units

sigma^2 estimated as 5806
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-1.024093	75.87867	59.72938	-0.2612447	2.158747	0.2146559	0.06104095	NA
Test set	177.285681	318.68983	284.74861	7.0032947	12.767500	1.0233317	-0.16668111	0.721025

2.2 Partition the data using the period until December 1993 as the training period. Run a neural network using R's nnetar function with 11 non-seasonal lags (i.e., p = 11)



	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
X1	156	-1.024093	76.11612	-1.735196	-2.982938	-160.0371	278.507	438.5441	0.3847116	0.3394057

2.3 Plots **above** show original and predicted data on far left. Fit appears very good. Residuals mean appears zero and stable and distribution in QQ plot which appears normal. Tests do not allow rejection of zero mean hypothesis or lack of kurtosis, but

do allow rejection of hypothesis of symmetry at the 95% confidence level. The histogram (not shown) supports normal appearance. ACF and PACF show significant autocorrelation at 12th and 14th lag possibly indicating chance given these are 95% intervals or that residuals capture relationships that the neural net skipped over.

2.4 NN to forecast sales in 1994.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-0.2494936	76.19647	58.66948	-0.2289607	2.130043	0.2108468	-0.02166766	NA
Test set	174.0174376	319.55588	279.08025	6.1302602	12.386779	1.0029607	-0.03809735	0.6749868

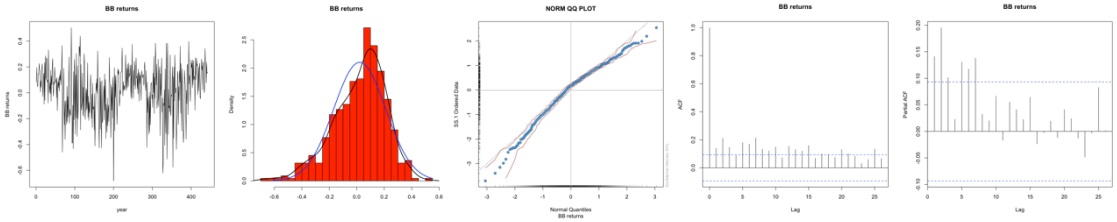
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	15.49844	285.23206	222.31979	-0.04620971	7.098462	0.7989745	0.06838359	NA
Test set	-57.51901	98.08848	79.45385	-5.23524620	6.634152	0.2855418	-0.50000000	0.05298271

2.5 Compare NN to ETS. ETS model selected is MAM (multiplicative error and seasonality and additive trend).

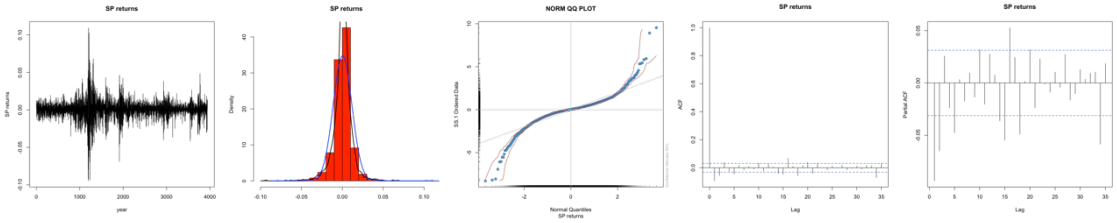
The first table **above** presents the NN model and the second the ETS model. NN appears to overfit the trainig data. ETS appears to generalize better than NN from training to test by MAPE measure and exceptionally so by RMSE where one could argue that the ETS underfits the training data. Interesting difference for MAPE vs RMSE for ETS. Possibly the range of data (the declining trend) has something to do with this as MAPE is percent of actual data. Possibly errors are more weighted toward earlier data when variance mean levels were greaster.

RMSE on all models assembled in Assignment 1 were in high 500s. So I am not sure what happened there. Favorite in that case was same as here ETS MAM.

3.0 HMM bull bear and S&P500



	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
bb return	442	0.02178872	0.1898210	0.05617107	0.03421163	-0.6812175	0.5032055	1.1844230	-0.64195713	0.4642184
bb sd	10	0.16978243	0.0545797	0.16619580	0.16909410	0.1008692	0.2442024	0.1433332	-0.01246591	-1.6617469



	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
sp return	3942	0.0002462719	0.011459120	0.0006766824	0.0005654694	-0.094695125	0.10957197	0.20426709	-0.373323	11.9718654
sp sd	10	0.0103047694	0.005318281	0.0095231378	0.0093414703	0.004927522	0.02338841	0.01846089	1.308683	0.8505071

3.1 EDA

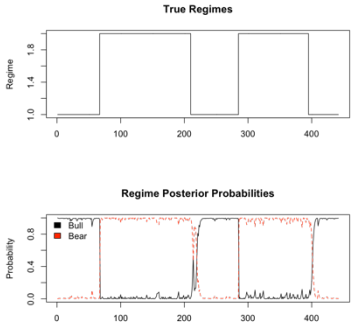
The above tables provide EDA for bullbear "bb" returns and S&P500 "sp" 1st differenced log returns and their respective standard deviations (2nd line in each table). The plots are for respective return series.

Tests of means, symmetry and kurtosis for "bb" allow rejections of zero means at the 95% confidence level (p-value = 0.01622), rejection of symmetry at the highest confidence level, rejection of no excess kurtosis at the 95% confidence level. Similar tests for "sp" do not allow rejection of zero mean hypothesis but do allow rejection of hypotheses regarding zero skew and kurtosis at the highest confidence level.

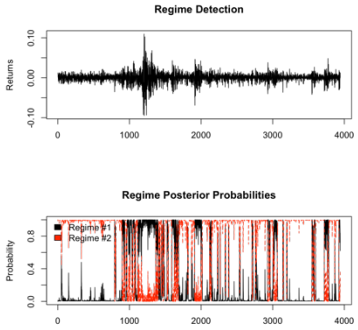
Time series plots of the 2 series are different. S&P500 after 1st differencing and logs show stationary zero mean with heteroscedasticity. BullBear returns show varying means and heteroscedasticity. One can see the moderate kurtosis and negative skew in the BullBear QQ but not so much in its histogram; most of the negative skew from outliers on the left that are also evident in the time series. Otherwise "bb" has a normal looking QQ plot. One can see the extreme leptokurtosis in the histogram and QQ plot and the timeseries seems to point to the source of this kurtosis in the extreme volatility left of center which may be the source of the outliers seen in the QQ plot. Also skew is evident in the placement of mean < median, and evident in the center of the distribution by trimmed < median.

BullBear PACF tails off exponentially while ACF is steadily above the interval top through lag 15 possibly representing MA or ARMA model. It would be difficult to interpret the ACF PACF for SP500 given the breaches on top and bottom. Possibly some form of sin relationship?

3.2-3.4

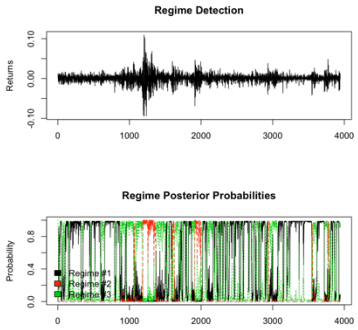


Bull bear true regimes and posterior probabilities appear to show the model is able to identify the true regime with slight lagged response.



The 2 state SP500 model **above** seems able to detect a shift in regime where regime 1 is a volatile state (observed in the range of x=1200, 1600, 1900, near 3000, 3700 and 3800) and regime 2 is stability. There doesn't seem to be much ambiguity except during the extreme volatile periods.

3.5 Fit 3 state S&P500



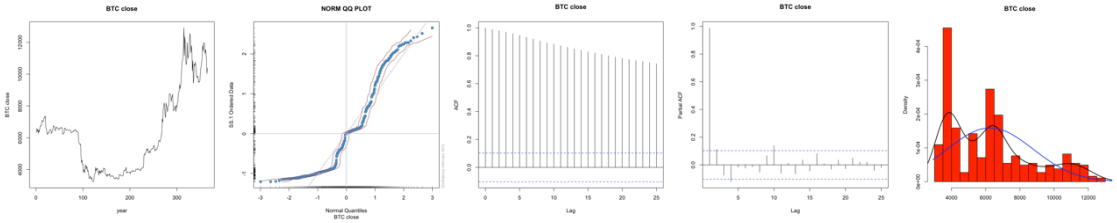
This model is more difficult to tell a story around. Switching between regime 1 and 3 sappears frequently when volatility is low. Regime 2 appears to indicate extreme volatility, only on 4 occasions. What is clear is that while switching between the 2 state model is more interpretable, there is more value in the switch between the state of (1 and 3) versus 2 in the 3 state model. In the 2 state model, volatility is coming lower after the

extreme levels near x=1200 and yet the HMM is still indicating regime 1. The 3 state model only retains regime 2 for the periods where volatility is extreme and very quickly adapts to regime 1 or 3 afterwards.

4. BTC to find exceptions.

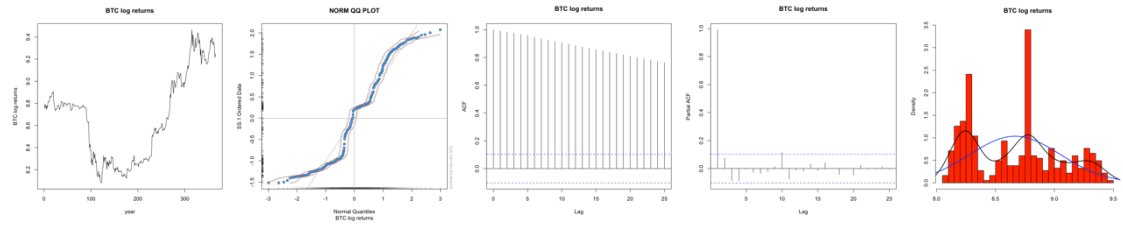
I would employ the 1st difference of log returns as the series demonstrates unit roots and variable means and serial correlations in other time series forms.

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
sp return	365	6258.5	2509	6267.3	5964.5	3233	12913	9681	0.7911	-0.3639
sp sd	10	639.2	518	459.9	588.1	128	1560	1432	0.4923	-1.4560



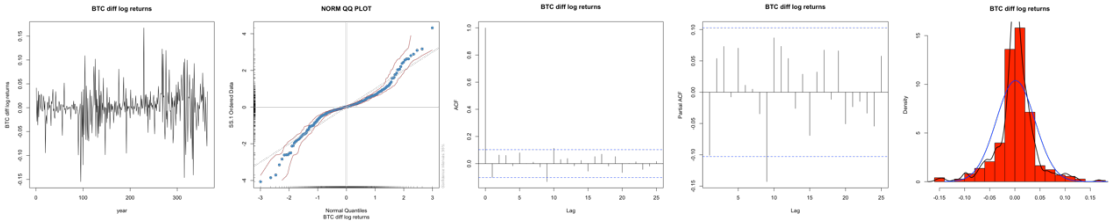
Bitcoin close natural form **above** shows variable means, positive skew and unit roots that are are verified by rejection of the null hypothesis of the KPSS test which indicates need for 1st differences. The second row of the table **above** shows the range in standard deviations of the series from 128 to 1560 and squares of the series exhibit unit roots as well.

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
BTC log returns	365	8.66611	0.38546	8.74310	8.64583	8.08101	9.4660	1.3850	0.2667	-1.090
BTC log returns sd	10	0.09867	0.06578	0.07796	0.09576	0.01976	0.2008	0.1811	0.3725	-1.631



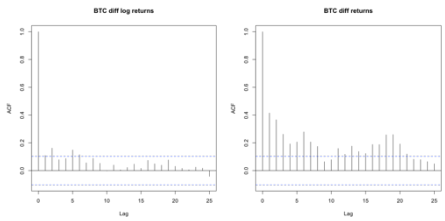
Similarly, **above** with log returns, there is evidence of unit roots in PACF at lag 1 and in ACF throughout. The mean varies in the timeseries plot on the left **above.** KPSS tests continue to show evidence for rejecting the null of no unit roots and suggest 1st differences.

	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis
BTC diff log returns	364	0.001361	0.03842	0.001922	0.002078	-0.15487	0.16722	0.32209	-0.1815	3.372
BTC diff log returns sd	10	0.036121	0.01227	0.038364	0.036412	0.01714	0.05278	0.03564	-0.1419	-1.726



1st differences of log returns provide stable time series mean which along with skew does not offer evidence to reject the null hypothesis of zero value. The distribution is kurtotic: we can reject the null with confidence at the highest level. KPSS rejects the null

hypothesis of unit roots and the ACF and PACF evidence no serial correlation. Generally the QQ plot shows a normal distribution.



Lastly, prefer diff log returns to diff returns as the square of the latter series show unit roots that are not apparent in the former when examining the ACF plots **above**. These results are borne out in the results of KPSS unit root tests of squared series that can be rejected for the series that is only 1st differenced.

Thus the following presents the results using the 1st differenced log returns of Bitcoin closes:

```
'data.frame':   227 obs. of  11 variables:
 $ Date     : chr  "2019-01-01" "2019-01-02" "2019-01-03" "2019-01-04" ...
 $ Open     : num  3747 3880 3961 3836 3874 ...
 $ High     : num  3939 3990 3966 3902 3927 ...
 $ Low      : num  3697 3826 3779 3784 3841 ...
 $ Close    : num  3880 3961 3836 3874 3855 ...
 $ Adj.Close: num  3880 3961 3836 3874 3855 ...
 $ Volume   : num  1.71e+08 2.11e+08 1.76e+08 1.71e+08 1.39e+08 ...
 $ year     : chr  "2019" "2019" "2019" "2019" ...
 $ mon      : chr  "01" "01" "01" "01" ...
 $ day      : chr  "01" "02" "03" "04" ...
 $ tdx      : num  20190101 20190102 20190103 20190104 20190105 ...

Wald's Sequential Probability Ratio Test (SPRT)

Decision:
Accept H1

Distribution: normal
n: 227, k: 1501575
h0: 0, h1: 1

Wald boundaries (log):
> B boundary:       -1.558
> A boundary:        2.773
> Likelihood ratio:  1501462

Preview k boundaries:
 n values     k      h0   h1
 1   3880  3880 -1.0581 3.27
 2   3961  7841 -0.5581 3.77
 3   3836 11677 -0.0581 4.27
 4   3874 15551  0.4419 4.77
 5   3855 19406  0.9419 5.27
```

5. There are some differences between the models of earlier this quarter versus models this week but most of the linear models from the 1st 9 weeks can be prepared with NN and RF models, just without the restrictions on the parameters. So we can fit ETS, AR and other models from earlier with NN and RF models of this week, but this weeks models provide much more flexibility to go beyond with additional hidden layers e.g. A trade off of course is that we cannot easily interpret NN models and the RF models do not provide predictive ability in the form of published parameters that we can hand over to customers.

Thank you for a great course. Im sorry I could not get all to you today as I am driving my kids to school tomorrow and had to stop and help them pack. I also had some issue with the software but all worked out. In any case, loved the course! Hope to see you sometime in CO.